APPROVED BY NTAPAA 3/22/05
Attachment C

NTAPAA Writing Assessment Discussion
Notes from 3/10/05 Meeting

Overall NTAPAA Feedback

The Kentucky Board of Education (KBE) and the Kentucky Department of Education (KDE) sought advice regarding the validity, reliability, and feasibility of specific changes to the writing assessment program recently approved by KBE. KDE also sought advice from the National Technical Advisory Panel on Assessment and Accountability (NTAPAA) regarding specific design decisions for various aspects of the writing assessment, such as rubric design and guidance for the writing portfolio audit.

NTAPAA members repeatedly emphasized the importance of clarity of purpose in their consideration of all the issues. Where the Board, KDE, or others desire to address multiple purposes, a policy decision must often be made about priority. For example, on the issue of whether the trait scores from the on-demand writing would be sufficiently reliable, NTAPAA essentially was split on reporting the student scores for consideration by teachers for improving instructional programs, but would not endorse the on-demand writing for student scores for high stakes individual student uses such as inclusion on a transcript or sole use to determine remediation. Thus, for their response to whether the proposed design of the on-demand writing assessment was "valid and reliable," NTAPAA requested the Board to distinguish clearly the purpose and use.

In the writing portfolio discussion, the panel members discussed some of the problems related to time and student ownership. Panel members encouraged development of an "administration manual" that addresses appropriate instructional strategies in development of potential portfolio entries. They encouraged KDE to include guidance regarding reasonable numbers of drafts, a definition of draft, and practices that diminish student ownership. Some members questioned the rationale for some of the proposed changes such as the reduction in the number of pieces, suggesting that a strengthened administration manual and a robust professional development plan would be more effective ways to address inappropriate and ineffective practice.

System Design

1. NTAPAA encouraged testing to be spread across grade levels, and to continue assessing a broader range of content areas than the No Child Left Behind (NCLB)-required reading and mathematics. The distribution of content areas by grades should consider many factors, including effect on school accountability, teacher workload, and alignment with significant grades in the curriculum. NTAPAA indicated that there is no research base and little consensus in policy across states about how much testing at a grade level is fair or appropriate within a high stakes accountability environment. Panel members encouraged the KBE to have a strong rationale for any policy
   decisions they reach concerning the spread of assessment across grade levels.

2. NTAPAA indicated that there is not a professionally established "standard for reliability and validity" that could be applied to the KBE Adopted Writing

Assessment Model. It suggested an appropriate strategy would be to address the reliability and validity of the writing assessment's individual parts and as a whole, and how the writing assessment fits into and impacts the entire assessment and accountability system.

Portfolio Review

1. NTAPAA emphasized that the sampling design of the portfolio review should reflect the purpose for the review. For example, if the primary purpose is to enhance credibility and a sense of fairness in the field, then more schools should be sampled each year. If the primary purpose is to adjust scores by school, then fewer schools should be sampled in greater depth. Some compromise between these two positions will probably be necessary, given the likely constraints of time and money. One promising option would be to design the review so that schools would have a high probability of being sampled over a biennium (e.g., 30-40% sample of schools, randomly sampled per year, with replacement; a modest number of purposefully selected schools could be included). The number of portfolios sampled per school would depend upon the desired statistical power to detect a likely event, e.g., a discrepancy of scoring of a certain amount, given a certain distribution of scores. This sampling approach would provide a strong estimate of writing portfolio performance and scoring accuracy for the state. Panel members also suggested that the KBE might consider an approach that targets review of only proficient and distinguished portfolios since past audits have indicated greater difficulty in scoring at those performance levels.

2. However, assuming the Portfolio Review would adjust portfolio scores of record, and thereby adjust school accountability scores, NTAPAA recommended that preferably all portfolios be rescored, rather than a sample. One way to do this, NTAPAA suggested, would be to execute a two-stage review, where schools that exceeded a certain level of discrepant scores would have all their student portfolios rescored. NTAPAA recognized that this recommendation may be operationally infeasible, and could not satisfy the Board's directive to review substantially more schools without significantly requiring more resources in time and money. Another option would be to adjust only the scores from the sample included in the review. A third option would be to recast the school's scores based on the review results; NTAPAA was less approving of this approach since it would adjust school scores based on an estimate rather than on observed student scores.

   NTAPAA members approved of the approach currently used where the school's original scores were considered in the review as the "first score" where the review provided a second score and discrepancies between the two scores were then resolved.

3. If KDE were to go to a regional scoring model for the portfolio review, rather than the centralized audit currently performed by the contractor, NTAPAA suggested that strong measures be taken to ensure as much as possible the scoring quality. These recommended measures included attending to similarity of scoring training across

regional scoring sites and instituting the types of scoring quality controls KDE has used in the past. Comparisons across regions and over time would strengthen the quality control.

Additional Performance Levels

1. NTAPAA thought that it made sense to create additional achievement levels for writing that in some way paralleled the sub-division of Novice and Apprentice in the other Kentucky Core Content Test (KCCT) subject areas. Regarding how these achievement levels should be defined and communicated, NTAPAA pointed out the differences between setting standards using a holistic single designation (as has been used in the writing portfolio in the past) and using multiple trait scores (as has been proposed for both the writing portfolio and the on-demand writing prompt). NTAPAA discussed a few possible approaches, including using a contrasting groups method based on profiles of the multiple traits, or using a body of work method based on a composite total score (weighted sum of scores on the multiple traits). In any case, NTAPAA urged KDE to do analyses and consult with content experts to determine how many traits could meaningfully be supported. NTAPAA, for example, guessed, based on other experience, that there may be as few as two major traits from a statistical view, but likely not six, although five or six trait scoring rubrics may reflect meaningful dimensions to writing content experts.

2. NTAPAA agreed that the current policy of using exact agreement or going to moderation to agreement was appropriate for a four-point rubric. NTAPAA also agreed that exact agreement was too stringent a scoring criterion for rubrics that involved more than four score points. Since KDE is considering a scoring rubric that will likely have more than four score points, NTAPAA agreed that it would likely be reasonable that KDE should move to a criterion of agreement-within-one-point (or more), but that this policy should be informed by the nature of the final rubric, the substantive meaning of the rubric score levels, and empirical studies. Empirical studies might include misclassification analysis, using, for example, total scores: if the cutscore for a Proficient classification were 11 points, a first score of 12.5 could have a second score discrepancy of 1.5 points without changing the student's classification. In that example, no moderation would be necessary if the two scores were 12.5 and 11.

On-Demand Assessment

1. NTAPAA recommended that KDE be sensitive to the unreliability of student scores based on a single writing prompt. NTAPAA noted that the test design should include multiple forms to ensure coverage across the three major modes of writing for school accountability and instructional feedback. The forms might be matrixed within a single year or over multiple years, preferably to ensure coverage within a biennium and to facilitate single year-to-year comparisions as desired. An alternate approach would be to equate prompts, which would reduce the statistical need for matrixing; multiple forms would be needed for instructional purposes. The number of forms should be informed by curricular considerations, not just statistical ones. Panel members provided guidance that if Kentucky wants all three modes of writing taught,

we should include all three in the assessment. The Office of Education Accountability (OEA) raised the possibility of having each student respond to more than one prompt using the same total time provided now, but NTAPAA did not discuss this option because it understood that KBE had decided against this approach or at least did not include this option in the adopted model.

2. In a lengthy discussion, NTAPAA strongly stated its reservations about using a single prompt on-demand writing assessment for any student stakes, including college placement. This position was based on the inherent unreliability of a one-item assessment, no matter how stringent the scoring. Thus, multiple-scoring would not make the performance more stable. NTAPAA members were divided about whether to provide overall scores and trait scores for instructional purposes, based on their views of likely misuse or benefits. All agreed that if such information were provided, it should be accompanied by strong qualifications. When questioned by KDE staff, NTAPAA members acknowledged that reporting and using on-demand writing scores based on a single prompt is fairly widespread both in other states' assessment programs and in higher education placement exams, and noted that they could not endorse such usage in these contexts either.

3. (See above for NTAPAA's advice on what extent the on-demand writing assessment can be used for individual student accountability or diagnostic feedback for individual students.)

Writing Skills Assessment (Grade 10)

1. NTAPAA again noted that a clear purpose for the writing skills assessment should determine its design, development, and use. If the purpose is to be a practice test for the ACT language arts subsection, then the writing skills assessment should mirror that as closely as possible. One NTAPAA member half-jokingly suggested using released ACT tests to construct the Kentucky test. NTAPAA did note that if the main purpose were to provide some prediction of success, then the state should do analyses of the predictive validity not only of the new assessment, but also of ACT and the college placement tests to actual performance of mechanics, grammar, etc. in postsecondary settings. That is, they felt that the ACT would not be a sufficient criterion. On the other hand, NTAPAA members emphasized that Kentucky's assessment program has been built up to now in identifying the valued content and skills through consideration of K-12 instruction, the breadth of coverage in the curriculum, and desired instructional targets. In that perspective, the state should identify the writing skills it desires students to possess by grade 10, and predictive links to ACT performance should be a secondary consideration.